УДК 504.064.2.001.18

Нейросетевой и геостатистический методы обработки экологической информации о распределении меди в верхнем слое почвы

Баглаева Е. М.[1,2*], Буевич А. Г.[1], Сергеев А. П.[1,2], Тарасов Д. А.[1,2], Арапов С. Ю.[2], Рахматова А. Ю.[2]

[1]Федеральное государственное бюджетное учреждение науки, институт промышленной экологии Уральского отделения Российской академии наук (ИПЭ УрО РАН), ул. С. Ковалевской, д.20, Екатеринбург, Россия, 620219
[2]Уральский федеральный университет, Мира, 32, Екатеринбург, Россия, 620002

Аннотация. Работа посвящена обработке экологической информации: предсказанию содержания меди в верхнем слое почвы по результатам почвенного скрининга. Для анализа использовались данные 101 пробы верхнего слоя почвы, отобранные с площади 6 км2 в городе Тарко-Сале, Ямало-Ненецкого автономного округа в 2007 году. Проведено сравнение оценок концентраций поверхностного распределения меди в почве, полученных с применением геостатистического метода (кригинг), метода искуственных нейронных сетей (ANN), а также гибридной модели (комбинирующих методы ANN и кригинг). После компьютерного моделирования была выбрана наилучшая структура ANN для восстановления поверхностного распределения меди. Сравнение содержания меди в почве, предсказанное методами кригинга и ANN, с гибридной моделью показали, что результат с наименьшей ошибкой предсказывает гибридная модель. Продемонстрированы возможности комбинации методов геостатистики и ANN для моделирования и анализа пространственно распределенных экологических данных.

Ключевые слова: обработка экологической информации, интерполяция, искусственные нейронные сети, кригинг, почва.

Neural network and geostatistical methods of processing the environmental information on the distribution of copper in the topsoil

Baglaeva E. M.[1]*, Buevich A. G.[1], Sergeev A. P.[1,2], Tarasov D. A.[1,2], Arapov S. Yu.[2], Rahmatova A. Yu.[2]

[1]Institute of Industrial Ecology UB RAS,
S. Kovalevskoy, 20, Ekaterinburg, Russia, 620990
[2]Ural Federal University, Mira, 32, Ekaterinburg, Russia, 620002

Abstract. The work deals with the processing of environmental information: the prediction of the copper content in the upper layer of the soil based on the soil screening. Data from the 101 samples of topsoil which was taken from the area of 6 km2 in Tarko-Sale, Yamalo-Nenets Autonomous district in 2007 was used. A comparison of the estimates of surface concentration distribution of copper in soil was obtained using a geostatistical method (kriging), artificial neural networks (ANN) and hybrid model (combining the methods of ANN and kriging). After computer modeling the best structure of ANN for the reconstruction of the surface distribution of copper was chosen. A comparison of copper content in the soil, predicted by kriging ANN and hybrid model, showed that the result with the smallest error was predicted by hybrid model. It was demonstrated the possibility of a combination of the methods of geostatistics and ANN for modeling and analysis of spatially distributed ecological data.

Keywords: processing of environmental information, interpolation, artificial neural networks, kriging, soil.

Введение

Одной из острых современных проблем в экологии является вопрос обработки данных экологического мониторинга и получения эффективных знаний об окружающей среде. Под экологическим мониторингом понимают комплексную систему наблюдений, оценки и прогноза изменений состояния окружающей природной среды [1]. В прошлом веке основными задачами работы с экологической информацией были получение данных наблюдений состояния компонентов окружающей природной среды, их накопление, хранение, структуризация, большей частью они были решены с развитием вычислительной техники, баз данных и систем их управления. Сегодня на первое место выходит проблема извлечения

(поиска) полезных знаний в базах данных. Эти знания могут быть представлены в виде закономерностей, правил, прогнозов, связей между элементами данных и др. Главным инструментом поиска знаний являются аналитические технологии DataMining, реализующие задачи классификации, кластеризации, регрессии, прогнозирования, предсказания и т. д.

Современный анализ экологических данных использует методы и технологии DataMining и включает использование детерминированных и/или геостатистических методов и построение моделей [2], решение задач классификации и регрессии, кластеризации и прогнозирования, а также интерпретацию и визуализацию результатов анализа.

Среди геостатистических методов интерполяции наиболее широко используется кригинг [3]. В экологических исследованиях кригинг показал значительные преимущества в прогнозировании загрязнения почвы, по сравнению с детерминированными методами интерполяции [4, 5, 6]. Тем не менее, результативность применения кригинга зависит от фактической пространственной неравномерности распределения моделируемых загрязнений, что в условиях гетерогенной среды делает применение этого метода интерполяции неэффективной.

В последние годы широкое распространение получили модели на основе искусственных нейронных сетей (ANN — ArtificialNeuralNetworks). Все большее количество исследователей применяют ANN в экологии, в том числе и в тех областях, где ранее использовали геостатистические методы. Так в [7] ANN применены для оценки концентрации озона в воздушном бассейне Сиднея, Австралия. Модель обеспечивает более надежные результаты оценки и предлагает более точные прогнозы концентрации озона. ANN были широко адаптированы и применены на практике исследователями [8] в свете возрастающих опасений по поводу экологических проблем, таких как глобальное потепление, частые явления Эль-Ниньо и аномалии циркуляции атмосферы. Методология, основанная на ANN была применена для экологического планирования, моделирования и получения высококачественных цифровых карт почв на земле Рейнланд-Пфальц (Германия), площадью около 600 км2 [9]. Авторы показывают, что подобный подход является экономически эффективным и обеспечивает надежные результаты. ANN использовались для прогнозирования долговременных изменений свойств почв и развитием таких процессов как деградация или опустынивание, являющихся одними из самых важных задач дистанционного зондирования. В [10] представлена методика скрининга данных и последующего применения ANN в области дистанционного зондирования.

Ряд исследований основан на применении гибридных моделей, включающих ANN. Так в [11] была выполнена оценка предсказательной эффективности четырех различных моделей; в частности, множественной линейной регрессии, однокомпонентной регрессии, искусственной нейронной сети (ANN) и сочетания однокомпонентной регрессии и искусственной нейронной сети для прогнозирования и для создания инструмента оценки концентрации мышьяка для Юго-Восточной Азии, включая Камбоджу, Лаос и Таиланд. Результаты моделирования показывают, что среди четырех различных моделей точность предсказания последней наилучшая. Исследование [12] было проведено с целью разработки моделей

сорбции в зависимости от основных свойств почвы с использованием ANN. В исследовании использовались данные по почвам, собранные на 133 сельскохозяйственных участках по всей Германии. Результаты сравнивались с данными, полученными на основе множественной линейной регрессии (MLR — MultipleLinearRegression). Характеристики моделей оценивали по среднеквадратичной ошибке, средней ошибке и эффективности моделирования. Было показано, что эффективность ANN в целом лучше показателей MLR.

В [13] производительность модели нейронной сети и множественной линейной регрессии оценивали с использованием набора тестовых данных. Результаты показали, что искусственная нейронная сеть с двумя нейронами в скрытом слое показала более высокую производительность в прогнозировании свойств почвы.

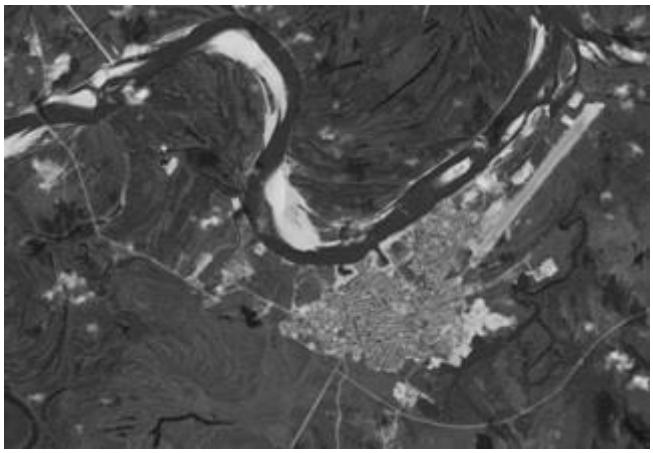
В некоторых работах сравниваются различные типы ANN и алгоритмов их обучения. Так в [14] сравнивается производительность двух подходов ANN, многослойный перцептрон (MLP — MultilayerPerceptron) и самоорганизующиеся карты Кохонена на примере цифровой картографии почв в Португалии и Испании. Авторы показывают, что лучшее исполнение ANN получается с моделью MLP, независимо от преобразования данных и метода отбора проб. Работа [15] посвящена сравнению традиционных методов обучения ANN: обратного распространения, Левенберга–Маквардта, квази-Ньютона, генетических алгоритмов и т. д. Результаты эксперимента показывают, что новый алгоритм «дифференциальная эволюция», относящийся к классу стохастических алгоритмов оптимизации и использующий некоторые идеи генетических алгоритмов, во многих практических случаях имеет более высокую точность и лучшую производительность, чем традиционные алгоритмы обучения.

В данной работе рассмотрены возможности применения для предсказания содержания химических элементов в верхнем слое почвы по результатам почвенного скрининга разных подходов: геостатистические методы (кригинг), метод ANN, а также гибридная модель, использующая ANN и кригинг.

1. Методы

Данные для настоящего исследования были получены по результатам почвенной съемки в городе Тарко-Сале, Ямало-Ненецкого автономного округа в 2007 году. Площадь, на которой были собраны пробы составила около 6 км2. Всего была отобрана 101 проба, пригодная для данного исследования. Долгота и широта для каждой точки отбора проб получены по данным GPS (global positioning system).

а)



б)

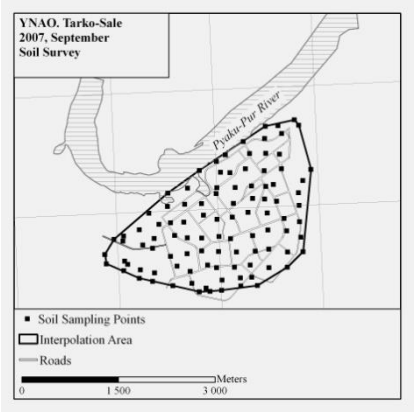


Рис. 1. Место почвенной съемки: а) г. Тарко-Сале (GoogleEarth, 2015); б) схема отбора проб

Физико-химический анализ проводился аккредитованной лабораторией по стандартным методикам. Для предварительного анализа характера распределения содержания меди в верхнем слое почвы были вычислены характеристики описательной статистики, построены диаграммы рассеяния и гистограммы распределения в пакете прикладных программ Statistica.

Для предсказания и визуализации картины содержания меди в верхнем слое почвы была проведена геостатистическая интерполяция (кригинг), использующая статистические особенности исследуемых точек вместе с пространственной автокорреляцией между ними и учитывающая пространственную конфигурацию точек отбора проб в исследуемой области. Существует несколько типов кригинга, включая простой кригинг, ординарный, универсальный, индикаторный и много других. Наиболее обще применимым является ординарный кригинг (OK — Ordinary Kriging). В обычном кригинге среднее значение считается постоян-

ным, но оно неизвестно. Кроме того, обычный кригинг при использовании локальной оценки не требует постоянства среднего по всей зоне оценивания; предполагается, что среднее постоянно только в окрестности точки оценивания. Число данных, использующихся при оценке, и значения весовых коэффициентов могут меняться в зависимости от местоположения оцениваемой точки x. данные выбираются из некоторой окрестности точки оценивания. Размер и форма этой окрестности зависят от исходных данных: предлагается использовать зону, ориентированную в соответствии с эллипсом корреляции. Уменьшение окрестности позволяет получать более вариабельную (менее сглаженную) оценку. Веса ординарного кригинга получают из уравнения кригинга, используя вариограмму. Параметры вариограммы и эффекта самородка могут быть оценены по эмпирической вариограмме. Несмещенной оценкой вариограммной функции является половина среднеквадратического различия между значениями пар данных:

$$\gamma\left(h\right)=\frac{\sum_{i=1}^{N(h)}\left|z\left(x_i\right)-z\left(x_i+h\right)\right|^2}{2N\left(h\right)}$$

где γ(h) это значение вариограммы для лага длиной h; и N(h) это число пар проб для длины лага h; и z(xi) и z(xi+h) это значение для двух точек разделенных лагом h.Вариограмма характеризует степень различия данных в зависимости от расстояния между ними. Чем ближе значения данных (меньше разница между ними), тем больше значение вариограммы [16].

Для оценки концентрации меди в тренировочном наборе данных был использован многослойный персептрон (MLP) с прямым распространением сигнала и методом обучения Левенберга–Марквардта. В простейшем случае, персептрон состоит из входного слоя, одного скрытого слоя и выходного слоя. Правило обучения используется для настройки весов и смещений персептрона так, чтобы приблизить значение выхода к целевому значению. Как правило, передаточные функции всех нейронов в сети фиксированы, а веса являются параметрами сети и могут изменяться. Ошибка для конкретной конфигурации сети определяется путем прогона через сеть всех имеющихся наблюдений и сравнения реально выдаваемых выходных значений с желаемыми (целевыми) значениями. Все такие разности суммируются в функцию ошибок, значение которой и есть ошибка сети.

Программирование нейронной сети проводилось в среде MATLAB® с использованием GUI интерфейса. В нашем случае использовался персептрон, входным слоем которого являлись координаты точек отбора проб, скрытый слой состоял из нескольких нейронов и выходной слой представлял содержание меди в соответствующей пробе.

Предсказательную точность ANN можно улучшитьинтеграцией ординарного кригинга остатков и ANN. Начальной процедурой для кригинга остатков (ORK — Ordinary Residual Kriging) является предсказание нейронной сетью значений в тестовых точках. Остатки при работе нейронной сети можно определить как:

$$r(xi) = Z(xi) - Z_{ANN}(xi)$$

где $r(x_i)$ остатки набора данных x_i, $Z(x_i)$ — измеренные значения, $Z_{ANN}(x_i)$ — это значения, оцененные с помощью нейронной сети. Полученные остатки были оценены с помощь кригинга. Наиболее распространенный является ординарный (обычный) кригинг. Ординарный кригинг отличается от простого кригинга тем, что не предполагает знание среднего значения. В обычном кригинге среднее значение считается постоянным, но оно неизвестно. Кроме того, обычный кригинг при использовании локальной оценки не требует постоянства среднего по всей зоне оценивания; предполагается, что среднее постоянно только в окрестности точки оценивания.

Оценка обычного кригинга строится, как линейная комбинация исходных данных:

$r_{ok}(x)=\sum \lambda_i r(x_i)$

где $r_{ok}(x)$ — это оцененное значение в точке x с помощью ординарного кригинга, $\lambda_i(x)$ — это оптимальные веса с условием что $\sum\lambda_i = 1$, и $r(x_i)$ — это остатки нейронной сети для точки x_i

Для предсказания остатков области исследования было использовано приложение ArcGisc использованием ординарного кригинга.

Окончательная оценка содержания меди была получена как сумма оценки нейронной сети и оценки остатков с помощью кригинга.

$Y(xi)= Z_{ANN}(xi)+ rok(xi)$

Для оценки эффективности различных методов интерполяции были использованы два индекса, средняя абсолютная ошибка (MAE — MeanAbsoluteError) и среднеквадратичная ошибка (RMSE — RootMeanSquareError), которые вычислялись следующим образом:

$$MAE = \frac{\sum_{i=1}^{n} \left| x_{\mathrm{mod}\,i} - x_i \right|}{n};$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{\mathrm{mod}\,i} - x_i \right)^2}{n}};$$

где x_{modi} — предсказанная концентрация (ANN или кригинг), x_i — измеренная концентрация, n количество точек.

2. Результаты

Для всех почвенных проб был проведен физико-химический анализ, в концентрациях выше порога обнаружения были измерены следующие элементы: Cr, Mn, Ni, Co, Cu, Zn, Cd, Pb. Далее предсказывалось распределение содержания меди в верхнем слое почвы при помощи различных подходов.

Распределение концентрации меди было относительно равномерным на всей исследуемой территории. В табличной форме представлены результаты описательной статистики.

Таблица 1. Описательная статистика меди

Содержание элемента	Минимум, мг/кг	Максимум, мг/кг	Среднее, мг/кг	СКО *), мг/кг	Коэффициент вариации	Коэффициент ассиметрии	Коэффициент эксцесса	Медиана, мг/кг
Cu	3,57	48,8	15,0	6,30	0,42	2,03	10,4	13,4

*СКО — среднеквадратичное отклонение.

В настоящем исследовании сравнивалось несколько подходов к моделированию распределения концентраций химических элементов в верхнем слое почвы: геостатистические методы (OK), метод искусственных нейронных сетей MLP, а также гибридная модель, использующая многослойный персептрон и кригинг остатков (MLP-ORK). Весь набор данных был разделен на две группы: семьдесят процентов (70 проб) составили тренировочный набор для обучения нейронной сети, остальные пробы (31 проба) составили тестовый набор для проверки работы нейронной сети. Это разделение было выполнено случайным образом с помощью функции «create subset» в геостатистическом анализе для ArcGis 9.2. Тренировочный набор данных использовался для обучения нейронной сети и построения кригинга в ArcGis. Тестовый набор применялся для проверки точности предсказания как кригингом так и нейронной сетью.

Для выбора модели ANN использовали персептрон, входным слоем которого являлись координаты точек отбора проб, скрытый слой состоял из нескольких нейронов и выходной слой представлял содержание меди в соответствующей пробе. Выбор количества нейронов в скрытом слое проводилось по наименьшей среднеквадратической ошибке предсказания содержания меди в тренировочном (70 проб), тестовом (31 проба) и полном наборе данных (101 проба). Количество нейронов варьировалось от двух до двадцати. Каждая сеть обучалась по 500 раз и из них выбиралась наилучшая. Качество обучения сети проверялось коэффициенту корреляции и среднеквадратической ошибке между результатом работы сети и обучающим набором данных.

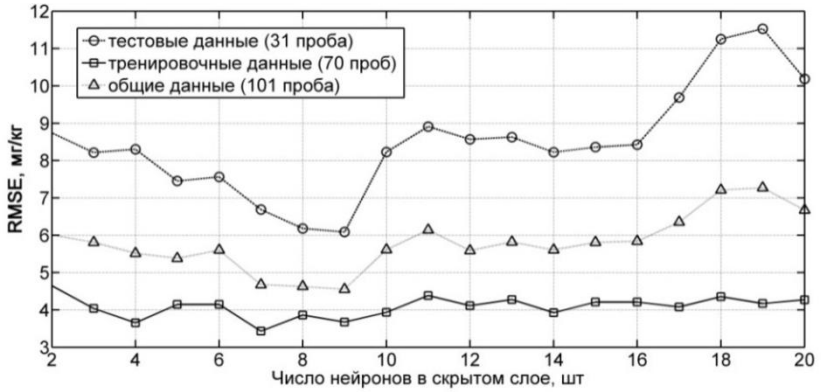


Рис. 2. Подбор количества нейронов в скрытом слое нейронной сети

Для моделирования распределения меди оптимальным количеством оказалось 9 нейронов в скрытом слое.

Таблица 2. Индексы оценки точности прогнозирования концентрации меди

Индекс	Средняя абсолютная ошибка (MAE), мг/кг			Среднеквадратичная ошибка (RMSE), мг/кг		
Метод	OK	MLP	MLP-ORK	OK	MLP	MLP-ORK
Cu	6,96	4,31	4,13	6,31	6,09	5,69

Сравнение методов показало превосходство ANN в точности моделирования. Также оказалось, что применение гибридного метода MLP-ORK дает увеличение точности прогноза распределения концентрации меди в верхнем слое почвы на 7 % относительно MLP и 11 % относительно кригинга, что согласуется с [17]. Оценка остатков ANN ординарным кригингом позволила сгладить высокие и скорректировать низкие значения концентраций меди в почве, что улучшило точность прогнозирования.

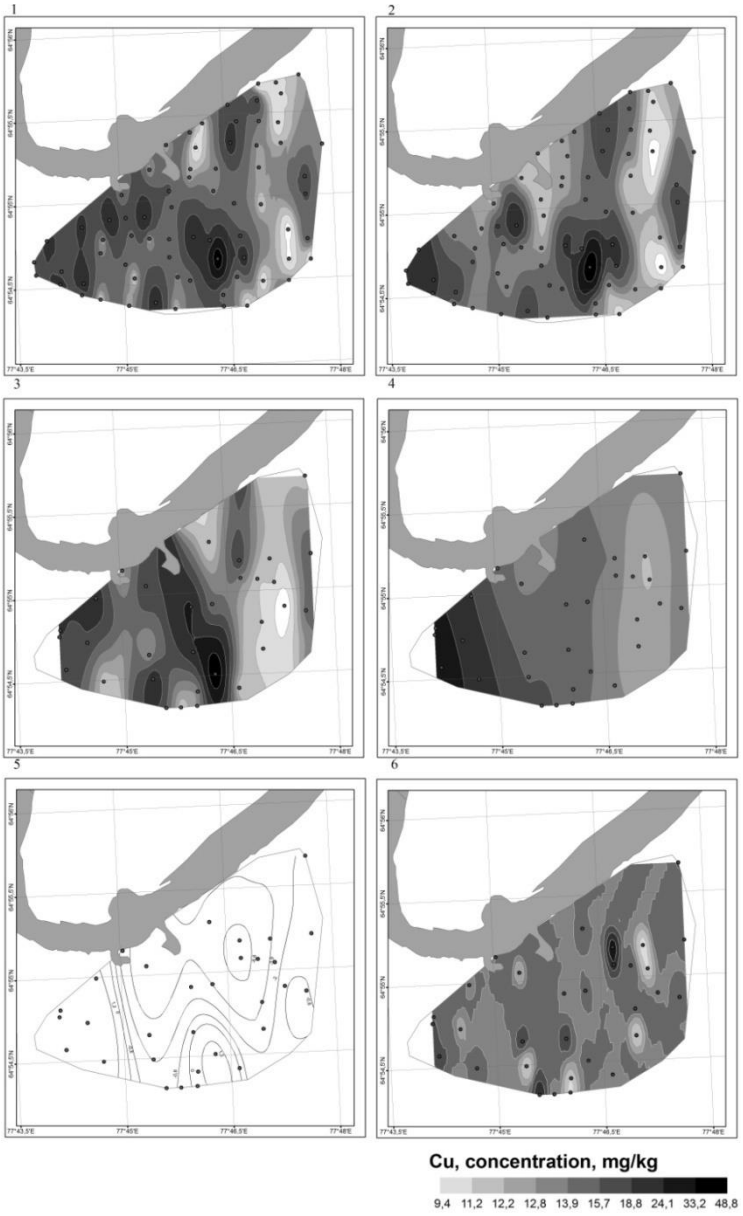


Рис. 3. Распределение содержания меди в почвенных пробах для различных моделей
(1. Кригинг по 70 тренировочным точкам; 2. Нейронная сеть по 70 тренировочным
точкам; 3. Кригинг по 31 тестовой точке; 4. Нейронная сеть по 31 тестовой точке; 5. Распределение остатков; 6. Нейронная сеть с кригингом остатков**)**

Выводы

Для обработки экологической информации, полученной в результате скринингов или мониторингов компонентов окружающей природной среды, предложено использовать комбинацию методов ANN и кригинга. Смоделировано распределение содержания относительно равномерно пространственно распределённого элемента в верхнем слое почвы на урбанизированной территории города Тарко-Сале, ЯНАО, Россия. Для моделирования использовалась искусственная нейронная сеть типа многослойный персептрон с прямым распространением сигнала и методом обучения Левенберга–Марквардта с двумя входными слоями, одним скрытым слоем и одним выходным слоем. Определено, что оптимальное количество нейронов в скрытом слое, дающее наименьшую среднеквадратическую ошибку предсказания для меди — 9. Результаты показали, что модель на основе многослойного персептрона оказалась точнее, чем модель на основе кригинга. Гибридный метод, включающий моделирование методом искусственных нейронных сетей и построение пространственного распределения остатков искусственных нейронных сетей ординарным кригингом, позволил еще уменьшить ошибку предсказания для меди.

Список литературы

1. Израэль Ю. А. Экология и контроль состояния природной среды. М.: Гидрометеоиздат, 1984. 560 с.
2. Анализ статистических зависимостей распределения загрязняющих веществ в поверхностном слое почвы урбанизированных территорий с применением математических моделей (LUR метод) / А. Г. Буевич [и др.] // Геоэкология, 2015, № 3. С. 268–279.
3. Webster R., Oliver M. Geostatistics for Environmental Scientists // John Wiley& Sons, Chichester. 2001. P. UL1–UL9.
4. Spatial variability of soil organic matter and nutrients in paddy fields at various scales in southeast China / Liu X. M., Zhao K. L., Xu J. M., Zhan M. H., Si B., Wang F. // Environ. Geol. 2008. № 53. P. 1139–1147.
5. Anomalies of chromium surface distribution in urban soils from subarctic region of Russia / Sergeev A. P., Baglaeva E. M., Antonov K. L., Medvedev A. N., Rakhmatova A. Y. // 15th International multidisciplinary scientific geoconference SGEM 2015. Water Resources. Forest, Marine and Ocean Ecosystems. Conference proceedings, V. II Soils, Forest Ecosystems, Marine and Ocean Ecosystems. 18-24 June, 2015, Bulgaria. P. 27–34.
6. Worsham L., Markewitz D., Nibbelink N. Incorporating spatial dependence into estimates of soil carbon contents under different land covers // Soil Sci. Soc.Am. 2010. J. 74, P. 635–646.
7. H. Wahid, Q. P. Ha, H. Duc, M. Azzi. Neural network-based meta-modelling approach // Applied Soft Computing. 2013. № 13. P. 4087–4096.
8. Liu ZeLin, Peng ChangHui, Xiang WenHua, Tian DaLun, Deng XiangWen, Zhao MeiFang. Application of artificial neural networks // Chinese Science Bulletin. 2010. № 34. P. 3853–3863.

9. Digital soil mapping using artificial neural network / Thorsten Behrens, Helga Forster, Thomas Scholten, Ulrich Steinrucken, Ernst-Dieter Spies, Michael Goldschmitt // Journal of Plant Nutrition and Soil Science. 2005. № 168. P. 1–13.
10. Remotely Sensed Soil Data Analysis / Filippo Amato, Josef Havel, Abd-Alla Gad, Ahmed Mohamed El-Zeiny // ISPRS Int. J. Geo-Inf. 2015. V. 4. P. 677–696.
11. Prediction of contamination potential / Kyung Hwa Cho, Suthipong Sthiannopkao, Yakov A. Pachepsky, Kyoung-Woong Kim, Joon Ha Kim // Water research. 2011. № 45. P. 5535–5544.
12. Estimation of heavy metal sorption / Ihuaku Anagu, Joachim Ingwersen, Jens Utermann, Thilo Streck // Geoderma. 2009. № 152. P. 104–112.
13. F. Sarmadian, R. Taghizadeh Mehrjardi. Modeling of Some Soil Properties Using Artificial Neural Network // Global Journal of Environmental Research. 2008. № 1. P. 30–35.
14. Sergio Freire, Ines Fonseca, Ricardo Brasil, Ricardo Brasil, Jose A. Tenedorio. Using Artificial Neural Networks for Digital Soil Mapping // AGILE 2013 – Leuven, May 14–17, 2013
15. Ngoc Tam Bui, Hiroshi Hasegawa. Training ANN Using Modification of Differential Evolution Algorithm // International J of Machine Learning and Computing. 2015. № 5(1). P. 1–6.
16. Демьянов В.В., Савельева Е.А.Геостатистика: теория и практика // под ред. Р.В. Арутюняна, Институт проблем безопасного развития атомной энергетики РАН: Наука, 2010. 327 С.
17. Fuqiang Dai, Qigang Zhou, Zhiqiang Lv, Xuemei Wang, Gangcai Liu. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau // Ecological Indicators. 2014. № 45. P 184–194.